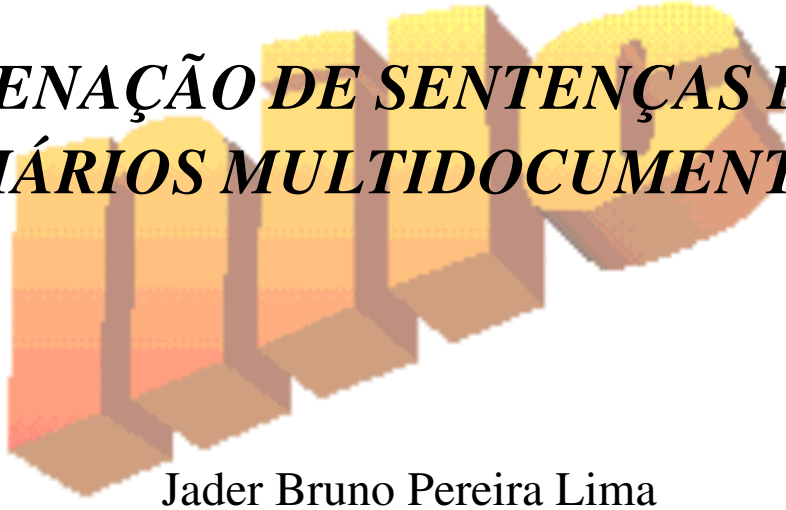
Universidade de São Paulo - USP

ORDENAÇÃO DE SENTENÇAS EM SUMÁRIOS MULTIDOCUMENTO

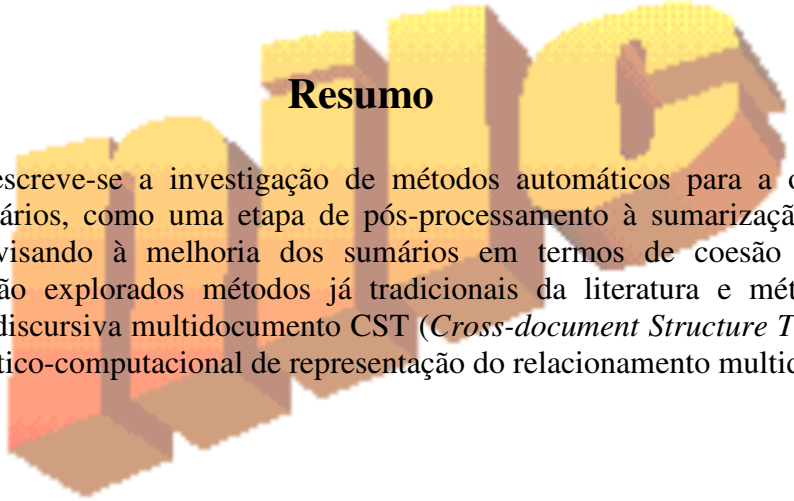


Jader Bruno Pereira Lima
Thiago Alexandre Salgueiro Pardo

NILC-TR-12-02

Junho, 2012

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil



Resumo

Neste trabalho, descreve-se a investigação de métodos automáticos para a ordenação de sentenças em sumários, como uma etapa de pós-processamento à sumarização automática multidocumento, visando à melhoria dos sumários em termos de coesão e coerência. Particularmente, são explorados métodos já tradicionais da literatura e métodos que se baseiem na teoria discursiva multidocumento CST (*Cross-document Structure Theory*), que é um modelo linguístico-computacional de representação do relacionamento multidocumento.

ÍNDICE

1. Introdução

A internet se tornou um vasto ambiente de comunicação, de onde podemos retirar facilmente todas as informações que possam ser relevantes para nossa vida. Segundo o IDC – *International Data Corporation*, em 2011 foram produzidos 1,8 trilhão gigabytes de dados na internet. Com isso, temos a nossa disposição inúmeras fontes de informação, criadas por vários autores e com várias abordagens para o mesmo assunto. Porém, essa quantidade de informações também pode ser um empecilho em tarefas de caráter mais ágil. Muitas pessoas utilizam a internet como principal fonte de informação das notícias do cotidiano, e é desejável que esta tarefa seja rápida e prática. Avaliando esse cenário, vemos que se faz necessário o desenvolvimento de aplicações computacionais para manipular esses dados, que estão disponíveis, em sua grande maioria, em forma textual.

Neste contexto, a sumarização multidocumento se torna uma tarefa importante no Processamento de Linguagem Natural, e sua função é produzir sumários, ou resumos, a partir de um conjunto de documentos sobre um mesmo tópico (Mani, 2001). Em geral, o processo principal desse tipo de sistema consiste em selecionar as sentenças mais importantes dos textos, as quais serão justapostas para formar o sumário. Esse tipo de sumário é chamado extrativo, pois é formado por sentenças extraídas integralmente dos textos.

A tarefa de sumarização possui três grandes processos: análise das características dos textos; transformação, com base na análise anterior, para a escolha das informações que estarão no sumário; e síntese, que engloba a composição e a apresentação do sumário. Os trabalhos deste projeto concentram-se nesta última etapa da criação de sumários multidocumento.

As sentenças de sumários gerados automaticamente necessitam de um processamento final antes de serem apresentadas ao usuário, e uma das principais tarefas pós-sumarização é a ordenação das sentenças do sumário, pois a ordem da narração dos fatos/eventos influencia diretamente na coerência e coesão dos sumários, principalmente neste cenário multidocumento, onde as sentenças provêm de diversos textos diferentes. Observa-se que, em sumários monodocumento, ou seja, sumários extraídos de um único texto, o problema da ordenação de sentenças

é praticamente inexistente, pois as sentenças escolhidas para compor o sumário devem seguir a mesma ordenação existente entre elas no texto original.

A seguir, na Figura 1, temos um sumário multidocumento em duas versões. Na primeira versão, suas sentenças estão dispostas sem nenhuma lógica de ordenação, já na segunda, as sentenças estão ordenadas de forma a ter maior legibilidade e coerência. Analisando os dois sumários da Figura 1, nota-se a importância da tarefa de ordenação de sentenças, e como esta produz uma melhora significativa na legibilidade do sumário.

Sumário sem Ordenação

Sentença 1: Os reatores nucleares foram desligados e não houve liberação de radiação.
Sentença 2: O Japão é um dos países do mundo mais suscetíveis a terremotos, com um tremor ocorrendo a ao menos cada cinco minutos.
Sentença 3: Chamas e rolos de fumaça preta foram vistos na usina nuclear de Kashiwazaki, que foi automaticamente fechada durante o terremoto.
Sentença 4: Um terremoto de 6,8 graus na escala Richter atingiu a costa noroeste do Japão nesta segunda-feira, 16, matando pelo menos sete pessoas na cidade de Kashiwazaki e deixando outros 700 feridos.

Sumário com Ordenação

Sentença 1: Um terremoto de 6,8 graus na escala Richter atingiu a costa noroeste do Japão nesta segunda-feira, 16, matando pelo menos sete pessoas na cidade de Kashiwazaki e deixando outros 700 feridos.
Sentença 2: Chamas e rolos de fumaça preta foram vistos na usina nuclear de Kashiwazaki, que foi automaticamente fechada durante o terremoto.
Sentença 3: Os reatores nucleares foram desligados e não houve liberação de radiação.
Sentença 4: O Japão é um dos países do mundo mais suscetíveis a terremotos, com um tremor ocorrendo a ao menos cada cinco minutos.

Figura 1: Sumários com e sem a ordenação de sentenças

Vários estudos surgiram nesta linha de pesquisa, os quais comprovam a necessidade de se criar métodos de ordenação para as sentenças. Por exemplo, para se ter uma ideia dos benefícios obtidos com tal tratamento, em um estudo realizado com juízes humanos (Barzilay et al., 2002), classificaram-se 10 sumários multidocumento gerados automaticamente, de acordo com a sua compreensão, em três níveis: compreensíveis, parcialmente compreensíveis ou incompreensíveis. A Tabela 1 mostra os resultados da classificação antes e depois das sentenças dos sumários serem ordenadas. Vemos que é de essencial importância esse tratamento das sentenças do sumário, para que possamos obter textos mais compreensíveis. Há uma melhoria considerável do número de textos que, antes da ordenação, eram considerados incompreensíveis.

Tabela 1: Classificação de sumários multidocumento

10 sumários	Sumários sem ordenação	Sumários com ordenação
Incompreensíveis	7	3
Parcialmente compreensíveis	2	2
Compreensíveis	1	5

Além da ordenação das sentenças, existem outras formas de pós-processamento de sumários. A fusão de sentenças, utilizada nos trabalhos de sumarização de Jorge e Pardo (2011) é uma técnica que pode melhorar a qualidade dos sumários. Chaves e Rino (2007) apresentam uma técnica de resolução de anáforas, que pode se utilizada no pós-processamento dos sumários, diminuindo a redundância das sentenças do sumário.

Neste relatório, apresentamos um estudo sobre ordenação de sentenças, com o desenvolvimento de métodos de ordenação, que são divididos em dois grupos: métodos simples de ordenação, que utilizam informações superficiais das sentenças como parâmetros de ordenação; e métodos informados de ordenação, que utilizam a teoria semântico-discursiva CST – Cross-document Structure Theory (Radev, 2000).

As relações CST relacionam pares de sentenças de textos distintos, atribuindo alguma informação a essa relação. Essas relações podem conter informação a respeito do conteúdo e da forma das duas sentenças relacionadas. Por exemplo, entre as sentenças “S1: O avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade.” e “S2: O avião explodiu e se incendiou, acrescentou o porta-voz da ONU em Kinshasa, Jean-Tobias Okala.” pode haver uma relação CST “Follow-up”, que nos diz que a sentença S2 versa sobre um fato ocorrido antes do fato narrado pela sentença S1. Mais detalhes sobre as relações CST estão contidos na Seção 2.2.

A seguir, na Seção 2, são apresentados alguns dos mais importantes trabalhos realizados na área de ordenação de sentenças, e também é apresentada a Teoria Semântico-Discursiva CST. Na Seção 3, são apresentados os métodos desenvolvidos neste trabalho, os quais estão divididos em dois grupos: os métodos simples e os métodos informados de ordenação de sentenças. A Seção 4 contém informações sobre o córpus utilizado para testes e também sobre as medidas exatas

de avaliação utilizadas neste trabalho, bem como as avaliações de todos os métodos de ordenação.

2. Revisão Bibliográfica

Na área de sumarização automática multidocumento, existem muitos trabalhos de ordenação de sentenças. A seguir, é mostrado um pequeno estudo de alguns deles, os quais são altamente relevantes para este trabalho.

2.1. Trabalhos Correlatos

Barzilay et al. (2002) nos mostram os desafios da sumarização multidocumento e propoem métodos baseados em duas caracteristicas bastante diferentes entre si. O primeiro método (Método da Ordenação por Maioria – OM) utiliza a informação topical existente entre as sentenças, onde as sentenças são classificadas em tópicos com o mesmo significado. Sendo assim, um grafo direcionado é construído, onde os nós representam os tópicos e as arestas representam a ordenação existente entre cada tópico. O peso de cada aresta nos indica o número de vezes que sentenças de determinado tópico aparecem antes das sentenças de outros tópicos. Na Figura 2 temos um exemplo deste grafo de precedência topical, construído a partir de três textos fonte hipotéticos. Por exemplo, a aresta de peso 2 que parte de Th_1 para Th_3 nos indica que há duas sentenças pertencentes ao tópico 1 que precedem sentenças do tópico 3 nos textos fonte. No canto superior da imagem, temos as ordenações dos tópicos nos três textos fonte diferentes, dado as sentenças que os compõem.

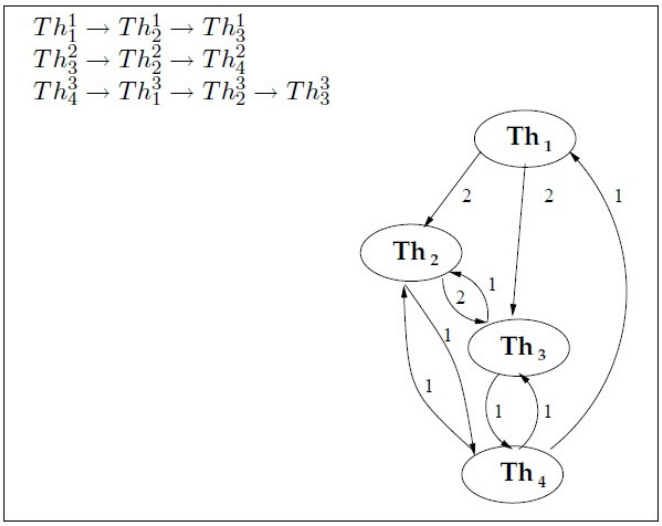


Figura 2: Construção do grafo de precedência topical

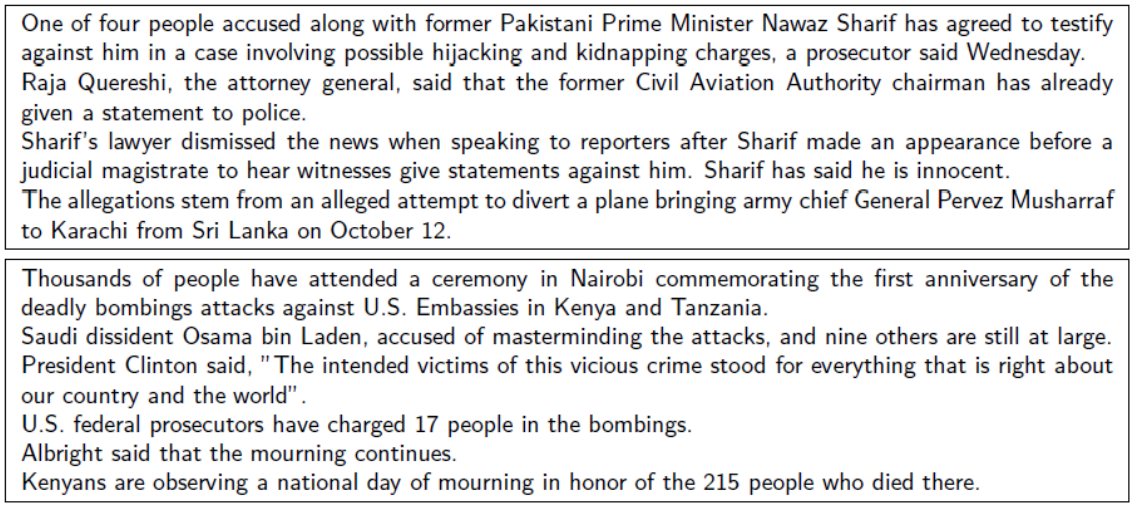
Avaliações mostraram que o método da Ordenação por Maioria se mostra bastante eficaz quando não há muita divergência na ordenação de sentenças entre tópicos, ou seja, todas as sentenças de um determinado tópico obedecem a mesma posição relativa às sentenças de outros tópicos. A seguir, na Figura 3, temos dois exemplos de sumários ordenados por este método, os quais foram avaliados por humanos. O primeiro é considerado um bom sumário, já o segundo é considerado um sumário razoável.

The man accused of firebombing two Manhattan subways in 1994 was convicted Thursday after the jury rejected the notion that the drug Prozac led him to commit the crimes.
He was found guilty of two counts of attempted murder, 14 counts of first-degree assault and two counts of criminal possession of a weapon.
In December 1994, Leary ignited firebombs on two Manhattan subway trains. The second blast injured 50 people – 16 seriously, including Leary.
Leary wanted to extort money from the Transit Authority.
The defense argued that Leary was not responsible for his actions because of "toxic psychosis" caused by the Prozac.

Hemingway, 69, died of natural causes in a Miami jail after being arrested for indecent exposure.
A book he wrote about his father, "Papa: A Personal Memoir," was published in 1976.
He was picked up last Wednesday after walking naked in Miami.
"He had a difficult life."
A transvestite who later had a sex-change operation, he suffered bouts of drinking, depression and drifting, according to acquaintances.
"It's not easy to be the son of a great man," Scott Donaldson, told Reuters.
At the time of his death, he lived in the Coconut Grove district where he was well-known to its Bohemian crowd.
He had been due to appear in court later that day on charges of indecent exposure and resisting arrest.
He sometimes went by the name of Gloria and wore women's clothes.
The cause of death was hypertension and cardiovascular disease.
Taken to the Miami-Dade Women's Detention Center, he was found dead in his cell early on Monday, spokeswoman Janelle Hall said.
He was booked into the women's jail because he had a sex-change operation, Hall added.

Figura 3: Sumários ordenados pelo método da Ordenação por Maioria

Os autores notaram que muitos sumários ordenados por este método que foram considerados razoáveis apresentavam problemas referentes à ordem cronológica das suas sentenças nos textos fonte. Com isso, outro método foi proposto (Método da Ordenação Cronológica - OC), o qual considera a data de publicação dos textos fonte para a ordenação das sentenças. As sentenças são agrupadas em tópicos com o mesmo significado, e a cada tópico é associado a data de publicação mais recente de suas sentenças. As sentenças do sumário são ordenadas pela ordem de seus tópicos, considerando que as sentenças pertencentes a tópicos com data de publicação mais recente devem ser ordenadas primeiro, e em caso de empate, ou seja, se duas sentenças forem pertencentes ao mesmo tópico, o desempate é feito considerando a posição das sentenças no texto fonte. A seguir, na Figura 4, temos dois exemplos de sumários ordenados com este método: o primeiro foi considerado um bom sumário por julgadores humanos, já o segundo foi considerado um sumário pobre.


One of four people accused along with former Pakistani Prime Minister Nawaz Sharif has agreed to testify against him in a case involving possible hijacking and kidnapping charges, a prosecutor said Wednesday.
Raja Quereshi, the attorney general, said that the former Civil Aviation Authority chairman has already given a statement to police.
Sharif's lawyer dismissed the news when speaking to reporters after Sharif made an appearance before a judicial magistrate to hear witnesses give statements against him. Sharif has said he is innocent.
The allegations stem from an alleged attempt to divert a plane bringing army chief General Pervez Musharraf to Karachi from Sri Lanka on October 12.

Thousands of people have attended a ceremony in Nairobi commemorating the first anniversary of the deadly bombings attacks against U.S. Embassies in Kenya and Tanzania.
Saudi dissident Osama bin Laden, accused of masterminding the attacks, and nine others are still at large.
President Clinton said, "The intended victims of this vicious crime stood for everything that is right about our country and the world".
U.S. federal prosecutors have charged 17 people in the bombings.
Albright said that the mourning continues.
Kenyans are observing a national day of mourning in honor of the 215 people who died there.

Figura 4: Sumários ordenados pelo método da Ordenação Cronológica (OC)

Estes dois métodos se mostraram parcialmente eficazes para certos estilos de textos fonte. O Método da Ordenação por Maioria (OM) se mostrou eficaz em textos fonte que possuem a mesma ordenação lógica das sentenças. Já o Método da Ordenação Cronológica foi bastante eficiente em textos baseados em eventos, porém os resultados são ruins em textos descritivos.

Buscando a melhoria dos resultados, os autores propuseram um terceiro método, que agrega características da Ordenação Cronológica e da Ordenação por Maioria. Este método, chamado de Método da Ordenação Aumentada, agrupa os

tópicos em grupos maiores, atribuindo uma relação de proximidade entre os tópicos de um mesmo grupo. Para cada um desses grupos, é atribuída uma data, que se refere a mais recente data de publicação do texto fonte de alguma sentença daquele grupo. Com isso, ordena-se cada um destes grupos por essa data, exatamente como é feito na Ordenação Cronológica. Dentro dos grupos, também é feita a ordenação de seus tópicos utilizando esse mesmo método OC. Sendo assim, o sumário tem as suas sentenças ordenadas, seguindo a ordem topical criada por este método. A seguir, na Figura 5, temos um exemplo de um sumário ordenado pelo Método da Ordenação Aumentada.

Thousands of people have attended a ceremony in Nairobi commemorating the first anniversary of the deadly bombings attacks against U.S. Embassies in Kenya and Tanzania. Kenyans are observing a national day of mourning in honor of the 215 people who died there.

Saudi dissident Osama bin Laden, accused of masterminding the attacks, and nine others are still at large. U.S. federal prosecutors have charged 17 people in the bombings.

President Clinton said, "The intended victims of this vicious crime stood for everything that is right about our country and the world". Albright said that the mourning continues.

Figura 5: Sumário ordenado pelo método da Ordenação Aumentada (OA)

Foram escolhidos 25 grupos de artigos para serem avaliados por juízes humanos. Os sumários poderiam ser avaliados como Bons, Razoáveis e Pobres, quanto a sua legibilidade, coerência e coesão textual. Os resultados estão na Tabela 2 a seguir. Note que os resultados do Método de Ordenação Aumentada (*Augmented Ordering*) foram bastante superiores aos demais (*Chronological Ordering* e *Majority Ordering*). Isso mostra que, no problema de ordenação de sentenças, não basta considerar apenas a ordem cronológica dos fatos, mas também a relação topical existente entre as sentenças.

Tabela 2: Avaliação dos métodos de ordenação (Barzilay et al., 2002)

	Poor	Fair	Good
Majority Ordering	3	14	8
Chronological Ordering	10	8	7
Augmented Ordering	3	8	14

Okazaki et al. (2004) nos mostram outra abordagem para o problema de ordenação de sentenças de sumários: utilizam relações entre sentenças, como se utiliza neste trabalho. Esta abordagem foi criada com a intenção de melhorar os resultados obtidos pela ordenação cronológica, que é bastante usada em vários trabalhos de ordenação multidocumento. O método proposto é dividido em três etapas.

Primeiramente, é feita uma segmentação topical entre os as sentenças escolhidas, ou seja, uma divisão das sentenças em grupos (*clusters*) com o mesmo tema/tópico. Na segunda etapa, após a segmentação topical, ordena-se cronologicamente, pela data de criação do artigo fonte, cada um desses clusters de sentenças. Depois disso, obtemos uma ordenação cronológica com separação topical. A fim de melhorar esta ordem, é realizada uma análise, nesta ordem cronológica obtida, de cada uma das sentenças com relação a suas sentenças antecessoras no texto fonte. Se a sentença não possui sentenças antecedentes ou só possui aquelas que são equivalentes em significado às ordenadas anteriormente, esta sentença é inserida em uma lista de ordenação. Caso esta sentença possua sentenças antecedentes com conteúdo ainda não ordenado anteriormente, é feita uma busca por essas sentenças para encontrar a sentença mais próxima, que não tenha nenhuma sentença antecedente.

Na avaliação do método de ordenação por precedência, os autores testaram seis métodos: Ordenação feita por Humano (HO), considerado o método que produz melhor resultado possível no trabalho; Ordenação Randômica (RO), considerado o método que produz o pior resultado possível no trabalho; Ordenação Cronológica (CO), ou seja, apenas a fase 2; Ordenação Cronológica com Segmentação Topical (COT) (fases 1 e 2); Método Proposto sem a Segmentação topical (PO) (fases 2 e 3); e o Método Proposto (POT).

Foi solicitado a humanos para avaliar a ordenação das sentenças nesses sumários. Essa avaliação se deu com classificações subjetivas em que cada juiz humano classificou a ordem das sentenças do sumário seguindo a escala a seguir: 4 (perfeito), 3 (aceitável), 2 (pobre) e 1 (inaceitável). Um sumário perfeito é um texto em que não se pode melhorar nada por meio de uma reordenação de suas sentenças. Um sumário aceitável é aquele que faz sentido e é desnecessária a revisão, embora haja espaço para melhorias em termos de legibilidade. Um sumário pobre é aquele que perde a discussão do fato em muitos lugares e requer uma alteração de menor importância para trazê-lo para um nível aceitável. Um sumário inaceitável é aquele que deixa muito a ser melhorado e requer uma reestruturação global, em vez de parcial. Os resultados estão exibidos em porcentagem na Tabela 3.

Podemos notar que os métodos que utilizam a relação de precedência das sentenças (PO e POT) são aqueles que conseguem as melhores classificações. Vemos também que em relação aos resultados dos métodos que utilizam apenas ordenação cronológica e segmentação topical, há uma significativa melhora na qualidade dos sumários.

Tabela 3: Classificação por especialistas humanos dos métodos apresentados

Método / Classificação	Perfeito	Aceitável	Pobre	Inaceitável
HO	0	0	6,0	**94,0**
RO	13,1	22,6	63,1	1,2
CO	10,7	22,6	61,9	4,8
COT	16,7	38,1	45,2	0
PO	15,5	36,9	44	3,6
POT	**52,4**	**21,4**	**26,2**	0

Lapata (2003) apresentou um método que ordena as sentenças levando em consideração a probabilidade condicional de uma sentença preceder outras sentenças nos textos fonte. Para o cálculo dessa probabilidade, inicialmente é necessário a identificação dos elementos que compõem as sentenças. Esses elementos são divididos em dois grupos: nomes e verbos, identificados com o auxílio de alguns córpora de treinamento. Também podem ser utilizadas, como elementos sentenciais, as dependências entre os nomes e os verbos, presentes na mesma sentença. A Figura 6 exemplifica as relações de dependência existentes entre elementos sentenciais. Como exemplo de relações verbais, no primeiro quadro da Figura 6, temos que o substantivo "*company*" tem uma relação com o verbo "*say*".

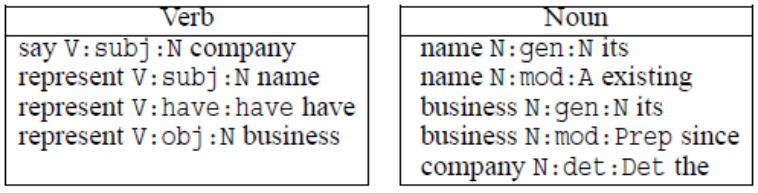


Figura 6: Dependências entre elementos sentenciais.

A probabilidade de uma sentença S1 aparecer antes de S2 é dada pelo produto das probabilidades de cada elemento de S1 aparecer antes de cada elemento de S2. A Figura 7 exemplifica um texto contendo três sentenças, e os relacionamentos existentes entre os elementos de cada par de sentenças adjacentes.

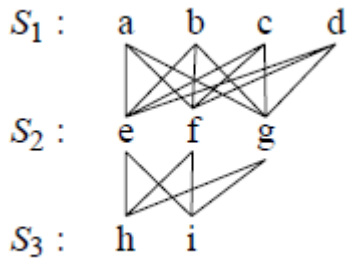


Figura 7: Exemplo de Probabilidade entre sentenças

A probabilidade de S3 aparecer depois de S2 (P(S3|S2)) é calculada pelo produto das probabilidades condicionais de seus elementos: P(h|e), P(h|f), P(h|g), P(i|e), P(i|f) e P(i|g), onde cada P(x|y) pode ser estimado pela Figura 6, dividindo a quantidade de arestas que conectam x e y pela quantidade de arestas que estão conectadas em x. No exemplo da Figura 6, P(h|e) = 1/6.

O algoritmo de ordenação proposto compara as probabilidades de cada sentença ocorrer antes de todas as sentenças restantes. Como esse problema é da classe dos problemas NP-Completos, foi proposto um método alternativo, com uma abordagem gulosa. Primeiramente, o algoritmo escolhe a sentença que deve estar no início da lista de ordenação, dado o conjunto de sentenças que se deseja ordenar. Essa escolha é feita comparando todas as sentenças de início em seus textos fonte: a sentença com maior probabilidade condicional, dada pela fórmula acima, é escolhida. A partir desta sentença, todas as probabilidades condicionais de todas as sentenças restantes aparecerem antes da sentença escolhida inicialmente são calculadas, e a próxima sentença da lista de ordenação é aquela que apresentar maior probabilidade de aparecer depois da sentença inicial. Toda sentença que é escolhida, é excluída da lista de todas as sentenças. A Figura 8 ilustra uma possível ordenação em um conjunto de três sentenças. A sentença inicial escolhida foi S2. A seguir, a probabilidade de S3 aparecer depois de S2 é maior que a probabilidade de S1 aparecer depois de S2, então, S3 é escolhida. Finalmente, S1 é escolhida, pois todas as outras foram excluídas do conjunto inicial. A ordem obtida pelas sentenças é S2, S3 e S1.

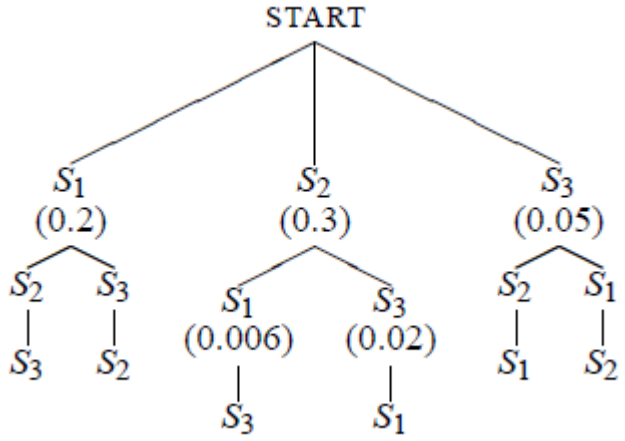


Figura 8: Exemplo de execução do algoritmo proposto

Para a avaliação do método, os autores utilizaram o coeficiente de Kendall, explicado com detalhes na Seção 4.1. Os testes foram feitos em 10 sumários multidocumento, extraídos de conjuntos de dois ou três textos fonte. Quatro variações do mesmo sumário foram produzidas: R, o qual tem as sentenças ordenadas randomicamente; H, o qual tem as sentenças ordenadas manualmente; NL, que possui suas sentenças ordenadas pelo algoritmo probabilístico descrito acima, utilizando apenas relações entre substantivos; e, por fim, VDND, que utiliza as dependências entre verbos e substantivos. Os resultados estão na Tabela 4, a seguir. A métrica utilizada para o avaliação foi o Coeficiente de Kendall (T), que é detalhadamente explicado na Seção 4.1.

Tabela 4: Avaliação de sumários ordenados pelos métodos propostos

Features	T	StdDev	Min	Max
B_R	.43	.13	.19	.97
N_L	.48	.16	.21	.86
$V_D N_D$	**.56**	.13	.32	.86
H_H	**.60**	.17	−1	.98

Nota-se que os resultados obtidos pelo método probabilístico VDND foram bem próximos dos resultados obtidos pela ordenação manual das sentenças do sumário, porém, este método tem um custo alto, pois a tarefa de identificação das dependências entre os elementos sentenciais é cara. Em contrapartida, o método probabilístico NL, que possui um custo computacional bem menor que o método VDND, atingiu resultados próximos ao método randômico. Isso nos indica que a identificação das dependências entre os elementos sentenciais é de suma

importância neste método probabilístico proposto pelos autores, quando aplicado à tarefa de ordenação de sumários multidocumento.

2.2. A Teoria Semântico-Discursiva CST

Os métodos de ordenação informados construídos neste trabalho são baseados na teoria linguística-computacional CST – Cross-document Structure Theory (Radev, 2000), e, a seguir, é apresentado um resumo desta teoria.

As relações CST conectam as sentenças dos textos fonte, definindo alguma informação existente entre elas. Essas informações se dividem em grupos (Maziero et al., 2010), como mostra a Figura 9.

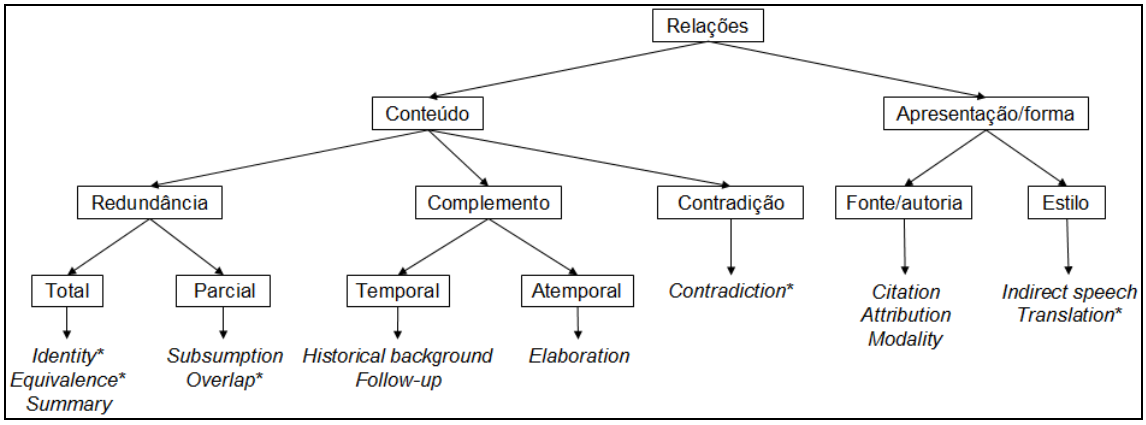


Figura 9: Tipologia das Relações CST

As relações se dividem em dois grupos: as que indicam alguma informação sobre o conteúdo das sentenças, e as que indicam informação sobre a apresentação das sentenças, principalmente.

O conteúdo das sentenças pode ser redundante (as duas sentenças apresentam a mesma informação), contraditório (há divergência na informação contida nas duas sentenças) ou complementar (uma sentença complementa de algum modo a informação contida em outra sentença). Já a apresentação das sentenças pode ter características relacionáveis em sua fonte ou em seu estilo de escrita. Frequentemente, pares de sentenças possuem pelo menos uma relação CST de forma e outra de conteúdo. A seguir, na Tabela 5, são definidos e exemplificados todos os tipos de relações CST (Aleixo e Pardo, 2008).

Tabela 5: As Relações CST

Relação CST	Definição	Exemplo
Identity	O mesmo texto aparece em mais de um local.	(S1) As vítimas do acidente foram 14 passageiros e três membros da tripulação. (S2) As vítimas do acidente foram 14 passageiros e três membros da tripulação.
Equivalence	Duas sentenças possuem a mesma informação contida.	(S1) O avião acidentado levava 14 passageiros e três tripulantes. (S2) As vítimas do acidente foram 14 passageiros e três membros da tripulação
Translation	Mesma informação contida em línguas diferentes.	(S1) Gritos de "Viva la revolucion!" ecoaram pela noite. (S2) Os rebeldes podiam ser ouvidos gritando "Viva a revolução!"
Subsumption	S1 contém toda a informação em S2, mais informação adicional que não está em S2.	(S1) Com três vitórias este ano, o Green Bay tem o melhor record na NFL. (S2) O Green Bay possui três vitórias este ano.
Contradiction	S1 e S2 apresentam informação conflitante.	(S1) A colisão no 26º andar ocorreu às 5:50 p.m. na quinta-feira, disse a jornalista Desideria Cavina. (S2) O avião colidiu no 25º andar do prédio Pirelli no centro de Milão.
Historical Background	S1 fornece contexto histórico da informação em S2.	(S1) Essa foi a quarta vez que um membro da família real se divorcia. (S2) Ontem o Duque de Windsor se divorciou da Duquesa de Windsor.
Citation	S1 explicitamente cita o documento S2.	(S1) O príncipe Albert continuou dizendo: "Eu nunca trapaceei". (S2) Um artigo anterior publicou o príncipe Albert dizendo: "Eu nunca trapaceei".
Modality	S1 apresenta uma versão mais qualificada da informação em S2, por exemplo, "é dito que; se sabe que".	(S1) Sean "Puffy" Combs é tido como um dos mais ricos. (S2) Puffy possui milhões de dólares em imóveis na área de Nova York
Attribution	S1 atribui a versão da informação em S2 usando, por exemplo, "de acordo com a CNN".	(S1) A colisão no 26º andar ocorreu às 5:50 p.m. na quinta-feira, disse a jornalista Desideria Cavina. (S2) O avião colidiu no 25º andar do prédio Pirelli no centro de Milão
Summary	S1 resume S2.	(S1) Os Mets ganharam o título em sete jogos. (S2) Depois dos exaustivos seis primeiros jogos, os Mets ganharam hoje novamente e levaram o título.
Follow-up	S1 apresenta informação adicional a qual tem acontecido desde S2.	(S1) 102 casualidades foram reportadas na área do terremoto. (S2) Até agora, nenhuma casualidade de abalo foi confirmada.
Indirect Speech	S1 indiretamente menciona algo o qual foi diretamente mencionado em S2	(S1) O presidente anunciou a população a garantia de novas moradias. (S2) "Eu garanto novas moradias", disse o presidente à população.
Elaboration	S1 informa detalhes de alguma informação dada mais generalizada em S2.	(S1) 50% dos estudantes estão abaixo de 25 anos; 20% estão entre 26 e 30 anos; o restante está acima de 30 anos. (S2) A maioria dos estudantes da universidade estão abaixo de 30 anos
Overlap	S1 informa fatos X e Y enquanto S2 informa fatos X e Z; Y e Z devem ser não triviais.	(S1) Hoje um pequeno avião bateu no 25º andar de um prédio no centro de Milão. (S2) Um pequeno avião bateu no prédio mais alto do centro de Milão na quintafeira à noite, expelindo fumaça dos andares mais altos.

Cada tipo de relação nos dá uma informação semântica sobre seu par de sentenças, e, para algumas dessas relações, é possível estabelecer uma ordenação relativa para a posição dessas sentenças no sumário, ou seja, qual das sentenças deve

aparecer primeiro no sumário, de modo que a coerência e a coesão deste seja a melhor possível. Mais detalhes de como é utilizada esta informação pode ser vista na Seção 3.2.

3. Métodos de Ordenação de Sentenças

Nesta seção são apresentados todos os métodos de ordenação de sentenças desenvolvidos, desde os mais simples, que utilizam informações básicas das sentenças, até os mais complexos, os quais fazem uso da teoria semântico-discursiva CST.

3.1. Métodos Simples de Ordenação

3.1.1. Método Baseado no Tamanho Sentencial

Partindo da hipótese de que o tamanho da sentença em palavras pode influenciar na ordenação do sumário, este método foi construído. O método simplesmente ordena os sumários, considerando apenas a quantidade de palavras de cada sentença. Duas versões deste método foram feitas, ordenando crescente ou decrescentemente as sentenças, de acordo com a sua quantidade de palavras. A seguir, na Figura 10, está um exemplo de um sumário ordenado com a versão deste método que ordena crescentemente as sentenças.

Sentença 1: Os reatores nucleares foram desligados e não houve liberação de radiação.
Sentença 2: O Japão é um dos países do mundo mais suscetíveis a terremotos, com um tremor ocorrendo a ao menos cada cinco minutos.
Sentença 3: Chamas e rolos de fumaça preta foram vistos na usina nuclear de Kashiwazaki, que foi automaticamente fechada durante o terremoto.
Sentença 4: Um terremoto de 6,8 graus na escala Richter atingiu a costa noroeste do Japão nesta segunda-feira, 16, matando pelo menos sete pessoas na cidade de Kashiwazaki e deixando outros 700 feridos.

Figura 10: Sumário ordenado pelo método do Tamanho Sentencial

Os resultados obtidos nestes dois métodos foram bem inferiores aos demais métodos deste trabalho, porém a ordenação crescente das sentenças obteve uma

ligeira vantagem em relação ao método de ordenação decrescente. Os resultados estão detalhados na Seção 4.2.

3.1.2. Método Baseado na Posição Textual

Em uma análise superficial de sumários ordenados manualmente, observamos que, em grande parte dos sumários, as sentenças mais próximas do início são as sentenças mais próximas do início de seu texto fonte. Utilizando esta informação, foi proposto um método de ordenação baseado na posição de cada sentença em seu texto fonte. O critério de desempate entre sentenças que possuem a mesma posição no texto fonte é o seu tamanho em palavras, onde as sentenças menores devem aparecer antes no sumário. A Figura 3 a seguir exemplifica e ilustra a ideia deste método.

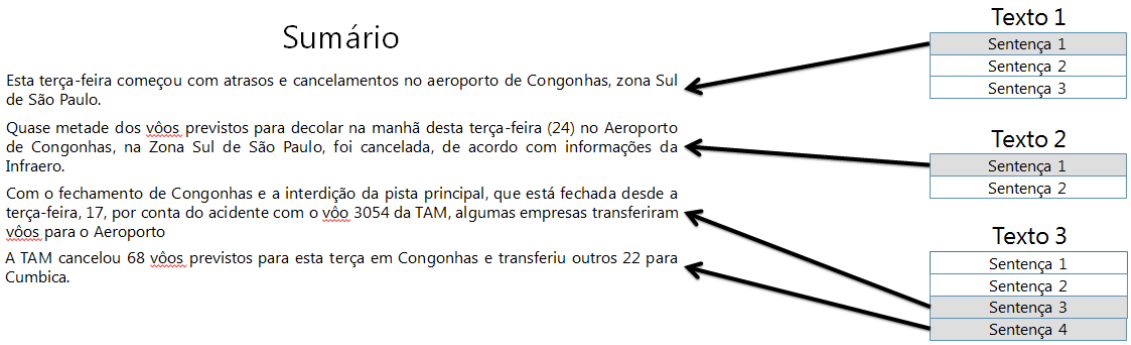


Figura 11: Sumário ordenado pelo método da Posição Textual

O sumário da figura anterior foi gerado a partir de três textos fonte. Pode-se notar que as duas primeiras sentenças do sumário foram escolhidas por serem a sentença inicial em seus respectivos textos fonte. O critério usado para o desempate entre elas é a quantidade de palavras que a sentença possui. Como a sentença vinda do Texto 1 é menor, em palavras, ela foi posicionada antes da sentença vinda do Texto 2. As duas próximas sentenças do sumário são provenientes do mesmo texto fonte, logo, pelo algoritmo, a sentença mais próxima do início do texto fonte é ordenada mais próxima do início do sumário. Os resultados obtidos na avaliação deste método foram os melhores deste trabalho e estão detalhados na Seção 4.2.

3.1.3. Método Baseado na Ordenação Topical

Além da posição textual e do tamanho sentencial em palavras, outro método de ordenação foi proposto, utilizando mais uma informação existente entre as sentenças dos textos fonte : a relação topical, via agrupamento das sentenças que abordam um mesmo assunto. O novo método proposto funciona da seguinte forma: calcula-se a similaridade entre as sentenças usando-se a medida de similaridade lexical do cosseno (Salton et al., 1997). Ela é computada entre cada sentença e as demais dos textos. A sentença é agrupada com o grupo de sentenças que resultou na maior similaridade (pois, teoricamente, abordam um mesmo tópico), desde que um limite mínimo de similaridade, calculado empiricamente, seja observado. Para este trabalho, consideramos esse limite com o valor de 0,2. Caso a sentença não se encaixe em nenhum grupo, é criado um novo grupo contendo essa sentença nova. A seguir, na Figura 12, temos dois textos fonte, e os tópicos identificados para esses textos.

Texto 1:
S1T1: TÓQUIO - Um terremoto de 6,8 graus na escala Richter atingiu a costa noroeste do Japão nesta segunda-feira, 16, matando pelo menos sete pessoas na cidade de Kashiwazaki e deixando outros 700 feridos.
S2T1: Equipes de resgate continuam procurando por pessoas em meio aos escombros.
S3T1: Chamas e rolos de fumaça preta foram vistos na usina nuclear de Kashiwazaki, que foi automaticamente fechada durante o terremoto.
S4T1: Segundo a TV NHK, o fogo atingiu um transformador de eletricidade e não houve vazamento de radiação.
S5T1: Uma série de pequenos abalos secundários atingiu a área, dentre eles um de 4,2

Texto 2:
S1T2: TÓQUIO - Um terremoto de 6,8 graus na escala Richter, com epicentro a 17 quilômetros de profundidade, atingiu a costa noroeste do Japão às 10h13m desta segunda-feira (22h13m de domingo em Brasília).
S2T2: O terremoto, que pôde ser sentido em Tóquio, foi seguido por outro tremor de menor magnitude, de 4,2 graus na escala Ritcher, às 10h34m (22h34m de domingo em Brasília).
S3T2: Um pequeno incêndio aconteceu em um transformador elétrico da usina nuclear de Kashiwazaki Kariwa, a maior do mundo, localizada perto do epicentro, mas o fogo já foi controlado.
S4T2: Os reatores nucleares foram desligados e não houve liberação de radiação.

Tópicos:
Tópico 1: S1T1 e S1T2; **Tópico 2:** S2T1; **Tópico 3:** S3T1 e S3T2
Tópico 4: S4T1 e S4T2; **Tópico 5:** S5T1 e S2T2

Figura 12: Tópicos identificados de dois textos fonte

Dados os grupos/tópicos criados, é feita uma ordenação entre eles, considerando a posição média de suas sentenças em seus textos fonte. Com os tópicos ordenados, cria-se uma lista com todas as sentenças de todos os textos, ordenadas crescentemente pela posição média. Como critério de desempate entre

sentenças de um mesmo tópico, são utilizadas duas informações sobre as sentenças: sua posição no texto fonte e seu tamanho em palavras, nesta ordem de prioridade. Com essa lista, temos a posição relativa entre as sentenças, e, então, ordena-se o sumário considerando as posições de suas sentenças nesta lista. A seguir, na Figura 13, temos um exemplo do funcionamento deste método.

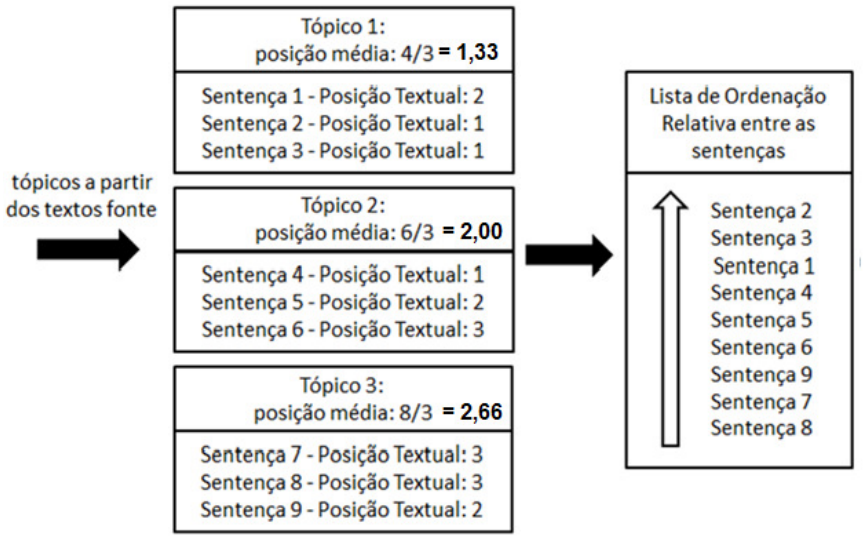


Figura 13: Ilustração do método da Ordenação Topical

Na figura anterior, temos a ordenação relativa de todas as sentenças presentes em três textos fonte. Pelo método de ordenação proposto, as sentenças foram agrupadas em três tópicos, dadas as medidas de similaridade existentes entre elas. Para cada tópico, é calculada a posição média de todas as suas sentenças, em seus textos fonte. Por exemplo, no Tópico 1, temos três sentenças, sendo que duas delas estão na posição inicial em seu texto fonte, e a outra ocupa a segunda posição. Fazendo a média desses valores obtemos 4/3, que é a posição média de todas as sentenças do Tópico 1. Esse procedimento é feito para todos os tópicos. A seguir, todos os tópicos são ordenados entre si, pelo valor crescente do valor médio de posição de suas sentenças em seus textos fonte, ou seja, as sentenças dos tópicos com menor valor médio de posição são ordenadas mais próximas do início da lista de ordenação. No exemplo as sentenças 1, 2 e 3 aparecem no topo da lista de ordenação, pois são pertencentes do tópico de menor valor médio de posições (Tópico 1). O método também ordena as sentenças dentro de cada tópico. Os critérios utilizados são: a posição textual da sentença e a sua quantidade de

palavras. No exemplo, o tópico 1 possui duas sentenças que possuem a mesma posição inicial em seu texto fonte, portanto, o método escolhe a menor sentença em palavras, e a ordena acima.

3.2. Métodos Avançados de Ordenação

Com o objetivo de se obter melhores resultados na ordenação das sentenças, foram desenvolvidos mais dois métodos, que utilizam as relações CST como base para heurística de ordenação das sentenças.

Partindo de um conjunto de textos que versam sobre um determinado assunto e com as relações CST bem definidas, temos um grafo (G), possivelmente desconexo, onde os vértices são as sentenças e as arestas são as relações CST entre cada par de sentenças. Neste grafo, são aplicados os dois métodos avançados de ordenação. A Figura 14 nos dá um exemplo hipotético do grafo G.

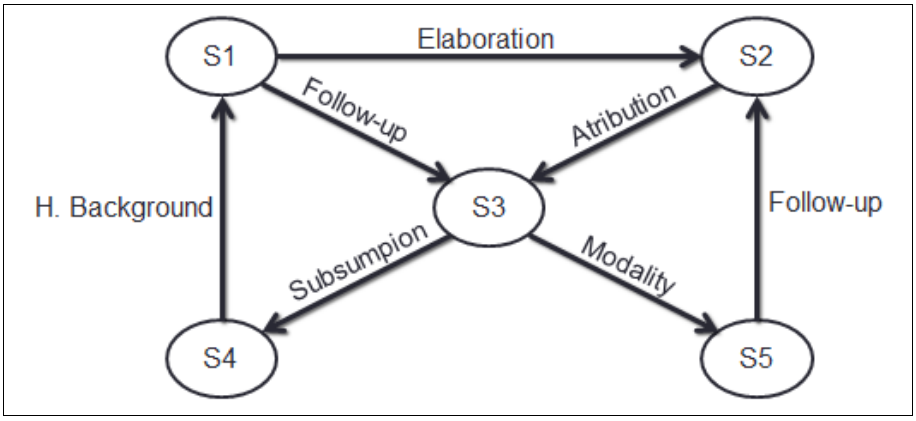


Figura 14: Exemplo de um grafo das relações CST.

Com as relações CST, temos informações que qualificam um par se sentenças, atribuindo a elas alguma característica semântica em comum. Por exemplo, na Figura 14, a relação *Historical Background* entre as sentenças S1 e S4 nos diz que a sentença S4 fornece um contexto histórico para a informação contida em S1.

Além dessa informação semântica, pode ser retirada uma informação extra das relações CST, que pode ser de grande utilidade para a ordenação das sentenças em um sumário. Essa informação extra se refere à ordem, dada a relação, que as sentenças deveriam ser apresentadas no sumário, de forma que a

coerência e a coesão textual sejam a melhor possível. Temos como exemplo disso a relação Follow-up na Figura 15 a seguir:

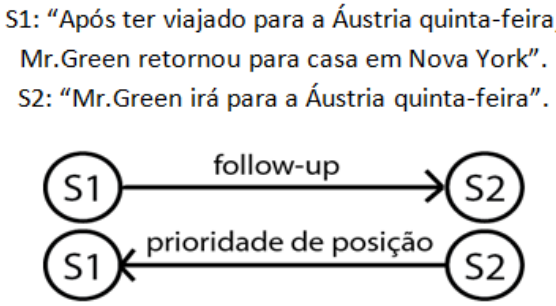


Figura 15: Exemplo de relação CST

Na Figura 15, podemos observar que a relação Follow-up, existente entre as duas sentenças S1 e S2, com uma ordem física definida de S1 para S2, nos diz que a sentença S1 apresenta informação adicional, a qual tem acontecido desde S2. Logo, em um sumário em que estão contidas as duas sentenças, vindas de dois textos fonte diferentes, a sentença S2 deve ser apresentada antes da sentença S1, ou seja, deve-se considerar a ordem lógica da relação CST, que, diferentemente da física, é definida de S2 para S1. Essa análise da ordem lógica deve ser feita para cada relação CST diferente, pois, em alguns casos, a ordem logica se mantem idêntica à ordem física da relação.

Temos a relação Subsumption (S1 -> S2) como outro exemplo, que nos indica que a sentença S1 contém toda a informação de S2, mais alguma informação adicional não presente em S2. Para garantirmos um fluxo natural de leitura do sumário, a sentença S2 deve ser apresentada antes da sentença S1. Isso se dá pelo fato de que, ao ler S1, o leitor do sumário já absorveria toda a informação de S2, futilizando o ato de ler a sentença S2, que simplesmente repetiria a informação já fornecida por S1. Com essa mesma análise, estabelecemos as ordenações relativas para outras relações, como Summary e Elaboration.

Assim, como pôde ser extraída uma informação de prioridade de apresentação da sentença na relação Follow-up anterior, podemos fazer essa mesma análise para cada relação CST, observando qual é a prioridade das

sentenças envolvidas na ordenação do sumário. Com isso, a partir do grafo das relações CST (G), gera-se outro grafo que nos indica a posição relativa entre as sentenças (G'), ou seja, a prioridade de posição no sumário entre as duas sentenças da relação. Cada relação CST tem a sua informação na ordenação de suas sentenças, como pode ser visto na Tabela 6, a seguir.

Tabela 6: Informação das Relações CST na ordenação das sentenças

Relação CST (S1 → S2)	S1 deve vir antes de S2	S2 deve vir antes de S1	Sem informação de ordenação
Atribution		x	
Citation	x		
Elaboration		x	
Follow-up		x	
Historical Background		x	
Indirect-speech	x		
Modality	x		
Subsumption		x	
Summary	x		
Contradiction			x
Overlap			x
Equivalence			x
Identity			x
Translation			x

Nota-se que algumas relações CST não nos dão nenhuma informação sobre a ordenação das sentenças, pois não possuem informação semântica para esse propósito. Por exemplo, as sentenças S1: *"A colisão no 26º andar ocorreu às 5:50 p.m. na quinta-feira, disse a jornalista Desideria Cavina."* e S2: *"O avião colidiu no 25º andar do prédio Pirelli no centro de Milão."* apresentam uma informação conflitante, porém não se pode definir uma prioridade de posição entre as duas. Essas relações não são mapeadas no grafo G', pois não possuem direcionalidade.

A partir desta análise, três métodos foram construídos, os quais utilizam o grafo de relações CST entre as sentenças para criar uma ordem relativa entre todas as sentenças de todos os textos fonte. Estes métodos são apresentados a seguir.

3.2.1. Método da Ordenação Topológica sem Análise Semântica das Relações CST

O primeiro método avançado consiste na ordenação topológica do grafo G', onde obtemos uma lista de vértices, ou sentenças, que nos indica a restrição de ordenação de cada uma das sentenças com relação às demais sentenças envolvidas na sumarização. Ou seja, uma sentença em uma posição k dessa lista deve ser posicionada acima das sentenças que estão na posição k+1, e também deve ser posicionada abaixo das sentenças que estão na posição k-1 da lista de ordenação topológica.

Devemos ainda observar que a ordenação topológica de um grafo só tem validade se este grafo for acíclico, e isto não pode ser garantido na anotação CST dos textos de onde são retiradas as sentenças do sumário. Neste método de ordenação, trataremos este problema de forma simples, apenas ignorando os ciclos do grafo, ou seja, não considerando vértices que já tenham sido visitados naquela iteração do método. Como parâmetro de sequência dos nós no algoritmo da ordenação topológica, são utilizadas duas informações sobre as sentenças (nós do grafo): a posição dela em seu respectivo texto fonte e a quantidade de palavras após a sentença no texto fonte. Para exemplificar a execução deste método, consideraremos um sumário extraído de dois textos fontes, contidos na Figura 16, a seguir. A Figura 17 apresenta as relações CST presentes entre esses dois textos, que serão usadas em nosso método de ordenação, já mapeadas em sua ordem lógica para o problema da ordenação de sentenças (grafo G').

Texto 1:
S1: SÃO PAULO - A pista principal do Aeroporto Internacional de São Paulo (Cumbica), em Guarulhos, será totalmente reformada em março de 2008, segundo informações do Ministério da Defesa anunciadas nesta segunda-feira, 6.
S2: Com isso, a reforma emergencial, que começaria em breve, foi descartada.
S3: O ministro da Defesa, Nelson Jobim, anunciou a reforma que, segundo estudos da Empresa Brasileira de Infra-Estrutura Aeroportuária (Infraero), a reforma poderá ser feita sem que a pista seja interditada.
S4: Apesar da definição, o cronograma da obra não foi divulgado.
S5: De acordo com informações da Defesa, a primeira etapa da reforma será feita com a reforma de um terço da pista, em uma das cabeceiras.
S6: Com isso, as outras duas partes ficam disponíveis para pousos e decolagens.
S7: Na segunda parte, a outra cabeceira será reformada e, na terceira etapa, o centro da pista será reformado.

Texto 2:
S1: RIO - O ministro da Defesa, Nelson Jobim, decidiu que será realizada uma reforma definitiva na pista principal de Guarulhos, o mais rápido possível, de acordo com a assessoria do ministério da Defesa.
S2: Fica afastada, portanto, a alternativa cogitada na semana passada, de se efetuar uma reforma emergencial nesse momento e se iniciar a reforma definitiva apenas em março de 2008.
S3: De acordo com estudos apresentados pela Infraero, será possível realizar a obra definitiva em três etapas, sem que seja necessário fechar a pista neste momento.
S4: O cronograma da obra depende de estudos finais que estão sendo realizados pela Infraero.

Figura 16: Dois textos jornalísticos a serem sumarizados

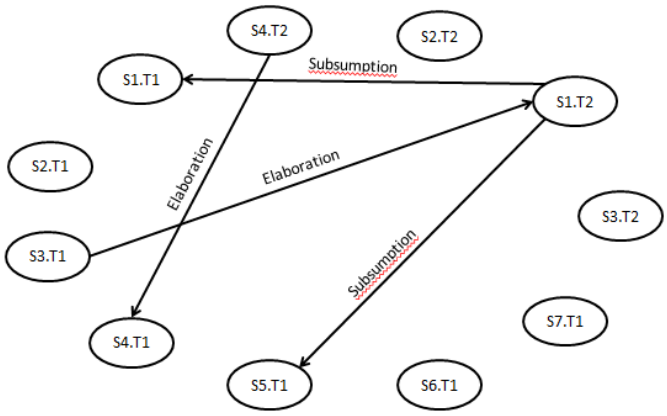


Figura 17: Grafo G' - Relações CST entre as sentenças dos textos da Figura X

O grafo G' da Figura 17 nos diz as restrições de posicionamento das sentenças no sumário. Por exemplo, entre a sentença 1 do texto 1 e a sentença 1 do texto 2 existe uma relação de Subsumption, que nos indica que a sentença 1 do texto 1 possui a mesma informação da sentença 1 do texto 2, porém adicionando mais dados informativos a ela. Se essas duas sentenças forem escolhidas por um sumarizador, logicamente, a sentença do texto 1 deve ter um posicionamento posterior a sentença do texto 2, pois possui um acréscimo de informação, aumentando o seu detalhamento.

Dado o grafo G', nosso método parte da sentença de menor posição e menor quantidade de palavras, no caso, a sentença 1 do texto 2. Pelo método da ordenação topológica, verificamos as sentenças adjacentes a S1T2 (sentença 1 do texto 2), sempre respeitando os critérios de escolha, que são, respectivamente, a posição da sentença no texto fonte e o seu tamanho em palavras. As sentenças adjacentes a S1T2 são S1T1 e S5T1, sempre respeitando a direcionalidade das relações. Seguindo nossos critérios, escolhemos a S1T1 para dar continuidade a ordenação topológica. Verificamos que a sentença S1T1 não possui relações CST, logo, não possui sentenças adjacentes, e então, essa sentença é retirada do grafo, e posta em uma lista, que nos indicará a posição relativa entre todas as sentenças dos textos. Seguindo o algoritmo da ordenação topológica, analisamos a outra sentença adjacente a S1T2, que é S5T1. Esta sentença também não possui sentenças adjacentes, então, assim como S1T1, é retirada do grafo e adicionada na lista de ordenação topológica. Como a sentença S1T2 não possui mais sentenças adjacentes, também é retirada do grafo e adicionada em nossa lista de ordenação

topológica. Prosseguindo com o algoritmo, verificamos qual é a sentença de menor posição contida no grafo. As duas sentenças S2T1 e S2T2 são candidatas a dar continuidade ao algoritmo, pois, com a exclusão de S1T1 e S1T2, se tornam as sentenças de menor posição em seus textos fontes. Escolhe-se a sentença S2T1, pois é a mais curta, em número de palavras. Como a sentença S2T1 não possui sentenças adjacentes, então é adicionada à lista de ordenação topológica do grafo. O método prossegue essa linha de execução, até passar por todas as sentenças de todos os textos fontes, formando uma lista de sentenças, ordenadas topologicamente, apresentada na Figura 18.

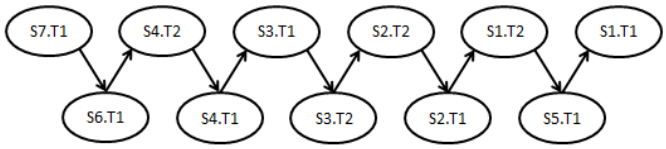


Figura 18: Ordenação Topológica do grafo da Figura 17

A Figura 19 nos mostra um sumário extraído desses dois textos e ordenados pela lista de ordenação topológica da Figura A. Notamos que o sumário produzido não é o ideal, pois a sentença S1T2 ficaria melhor posicionada no inicio do sumário, sendo complementada pelas outras duas sentenças que o compõem. Essa inadequação do sumário se deve ao fato de que o grafo CST anotado entre esses dois textos é muito esparso, e isso possibilita a criação de inúmeras listas de ordenação topológica diferentes. Por exemplo, a lista construída neste exemplo nos indica que a sentença S1T2 deve aparecer depois da sentença S7T1, porém, o grafo CST não nos impõe essa restrição.

Sumário:
S7T2: Na segunda parte, a outra cabeceira será reformada e, na terceira etapa, o centro da pista será reformado.
S1T2: RIO - O ministro da Defesa, Nelson Jobim, decidiu que será realizada uma reforma definitiva na pista principal de Guarulhos, o mais rápido possível, de acordo com a assessoria do ministério da Defesa.
S5T2: De acordo com informações da Defesa, a primeira etapa da reforma será feita com a reforma de um terço da pista, em uma das cabeceiras.

Figura 19: Sumário ordenado pelo método de ordenação topológica

Outra versão deste método, chamada Ordenação Topológica Invertida, foi desenvolvida, a fim de se melhorar os resultados obtidos. Esta segunda versão é

idêntica à primeira, diferindo da mesma em apenas um ponto. O algoritmo da ordenação topológica escolhe as sentenças mais distantes do início de seu texto fonte para prosseguir, diferentemente do método original, que escolhe as sentenças mais próximas do início. Este método foi desenvolvido com base nos resultados obtidos, e também por meio de análises nos sumários ordenados pelo método anterior.

Constatamos que a principal informação para a ordenação das sentenças é a posição de cada sentença em seu texto fonte. Porém, o algoritmo anterior prevalecia o inverso dessa heurística, pois as sentenças mais distantes do início tem a maior probabilidade de ficar no topo do sumário.

3.2.2. Método da Ordenação Topológica com Análise Semântica das Relações CST

O segundo método proposto neste trabalho tem como propósito a exclusão parametrizada dos ciclos do grafo G', ou seja, escolhendo-se as melhores arestas a serem excluídas. Essa tarefa é feita utilizando uma análise do tipo de relação CST que está contida na aresta. A Figura 20 ilustra os passos percorridos por esse método.

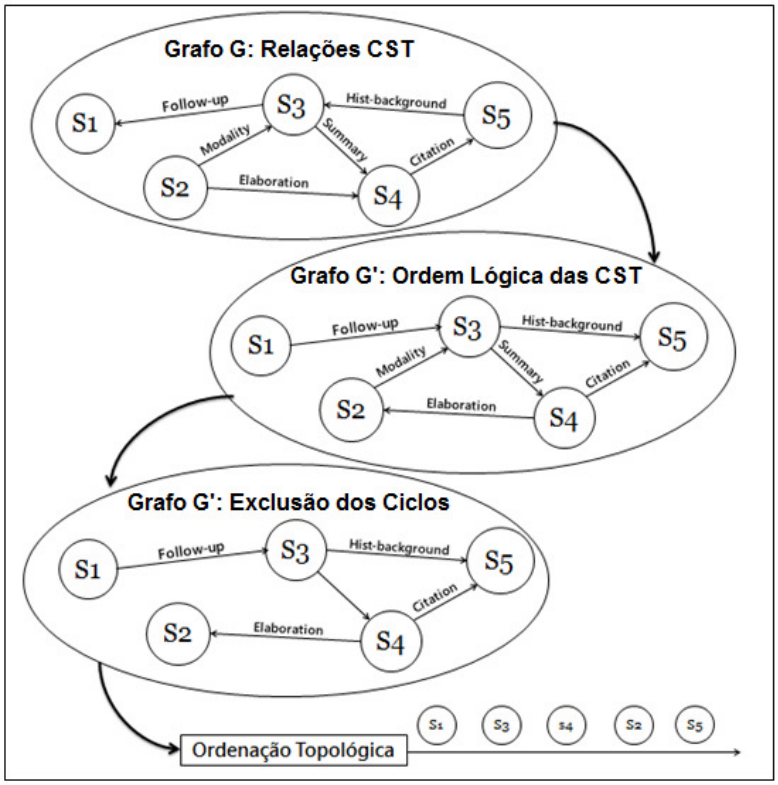


Figura 20: Ilustração do método da Ordenação Topológica sem Análise Semântica

Para a exclusão dos ciclos, avalia-se a semântica de cada relação CST, onde é feita uma ponderação das arestas de acordo com a força de informação que a relação CST nos dá entre a ordem relativa entre sentenças, ou seja, qual sentença deve vir antes. A Figura 21 nos mostra o peso de todas as relações CST, que são divididas em grupos de acordo com o tipo de informação cronológica que nos passam. Por exemplo, a relação Follow-up é mais pesada, pois nos dá maior certeza sobre a posição relativa entre as sentenças (um fato ocorrido antes, deve sempre aparecer antes no sumário). As relações marcadas com um asterisco não nos dão informação de ordenação, portanto seu peso é zero.

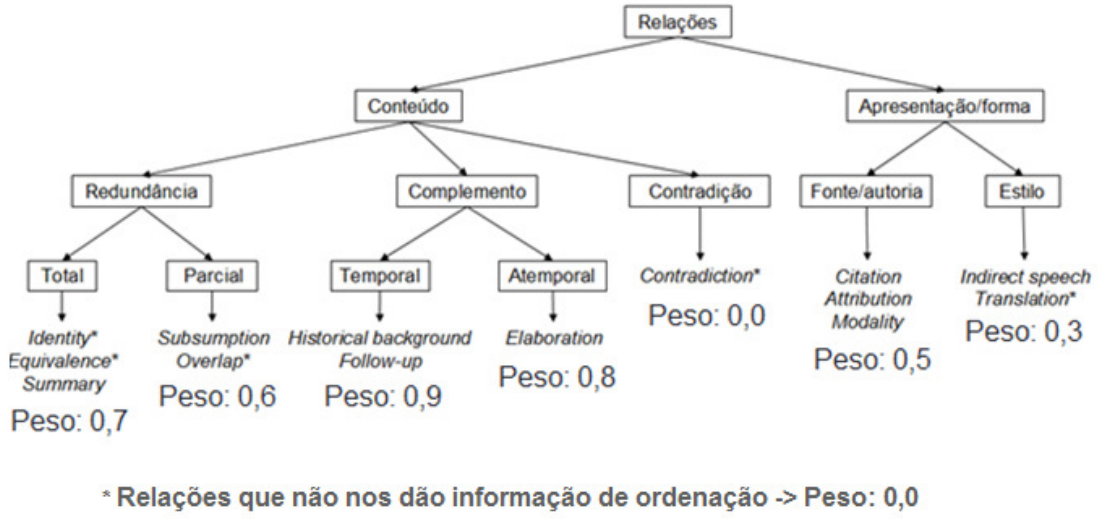


Figura 21: Ponderação das relações CST

Esses pesos foram determinados com uma análise superficial da importância de cada relação CST. As relações de Complemento nos dão uma informação totalmente baseada na ordenação das sentenças, onde cada sentença presente na relação tem uma informação crucial sobre a ordenação relativa entre elas. Por esse motivo, os pesos dados às relações deste tipo são maiores que os pesos dos demais tipos de relações CST. Diferenciando os grupos de relações restantes, temos as relações que indicam informação sobre o Conteúdo e sobre a Apresentação das sentenças. Para nossos métodos de ordenação, decidimos que as relações de conteúdo têm maior importância na tarefa de ordenação de sentença, e assim, é atribuído maior peso a essas relações de conteúdo.

Com o grafo todo ponderado, as arestas a se excluir do ciclo são escolhidas a partir de seu peso e também do número de ciclos de que ela faz parte, ou seja, quanto menor for o peso de sua relação CST e de mais ciclos ela fizer parte, ela é mais visada à exclusão. Essa abordagem foi pensada com a intuição de se excluir o menor número de arestas possível, mantendo a maior quantidade de informação semântica relevante para a ordenação. Essa informação é conseguida da seguinte forma: cada aresta contida em pelo menos um ciclo, tem seu peso dividido pelo número de ciclos que faz parte. A partir daí, escolhe-se a aresta com o menor desses valores para a exclusão. Assim como os método anterior, que não faz a análise semântica das relações CST, este método também possui a sua versão

invertida, que escolhe as sentenças mais distantes do início em seu texto fonte para prosseguir com o algoritmo da ordenação topológica.

Em todos os métodos que utilizam as relações CST, também foi adicionada uma relação de precedência entre cada sentença com relação a sua sentença anterior em seu próprio texto fonte. O peso dessa relação é 1,00, ou seja, maior do que os pesos de todas as relações CST, pois, observações superficiais mostraram que a posição relativa entre as sentenças de um mesmo texto é uma informação muito forte para a ordenação de sentenças.

Depois de excluídos todos os ciclos do grafo, a ordenação topológica é feita do mesmo modo como é feita no método da ordenação topológica sem análise das relações CST (vide Seção 3.2.1). A Figura 22 nos mostra a exclusão dos ciclos de um grafo G' hipotético.

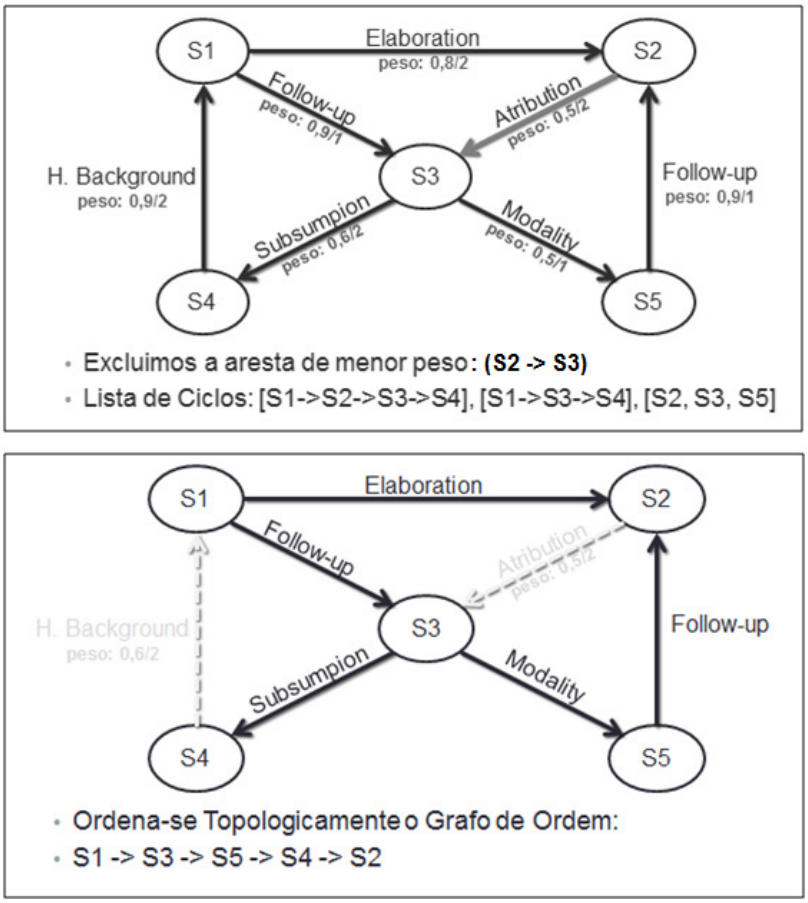


Figura 22: Exemplo de exclusão de ciclos no grafo G'

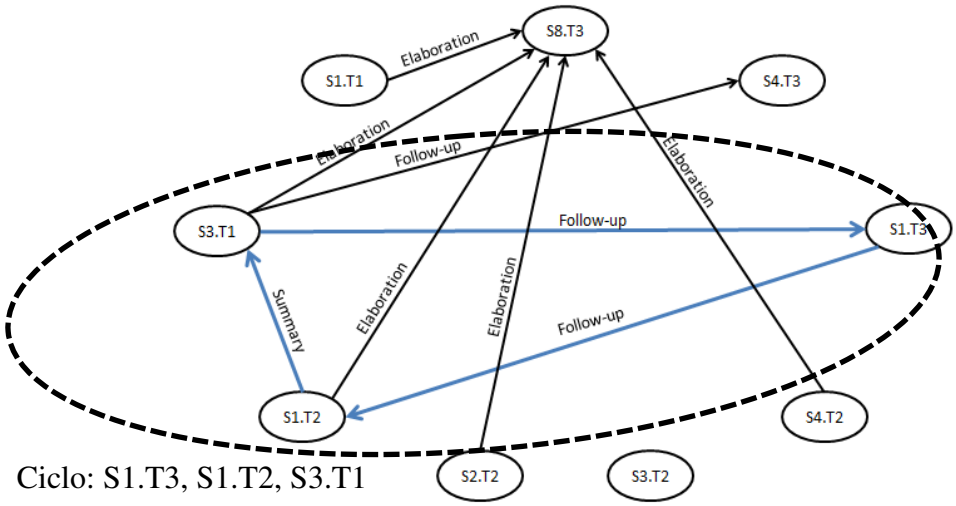
A seguir, na Figura 23, temos um conjunto de textos, que serão sumarizados por este método de ordenação topológica com análise das relações CST. A Figura 24 mostra as relações CST existentes entre todas as sentenças dos três textos fonte da Figura 23.

Texto 1:
S1: O furacão Dean mudou de rumo e não atingiu em cheio as ilhas Cayman nesta segunda-feira, e continua a avançar em direção à península de Yucatán, no México, depois de causar fortes chuvas e ventos na Jamaica e no Caribe, onde forçou a saída de milhares de turistas.
S2: Horas antes, havia forte receio de que Dean, um furacão de categoria 4 descrito por meteorologistas do Centro Nacional de Furacões (NHC, na sigla em inglês), com sede em Miami, atingisse em cheio as ilhas a uma velocidade de 240 km/h.
S3: Após causar fortes ventos e chuvas na Jamaica, furacão Dean segue para o México.
S4: "Seja lá em que Deus vocês acreditam, é hora de abaixar as cabeças e rezar para ele", disse o coordenador das equipes de retirada, Zemrie Thompson, às pessoas levadas para abrigos.
S5: O centro de Dean passou cerca de 160 km ao sul das ilhas Cayman, causando ventos de até 92 km/h.

Texto 2:
S1: KINGSTON - O furacão Dean chegou à costa sul da Jamaica, inundando a capital e espalhando árvores e telhados depois de matar nove pessoas na passagem pelo Caribe nesta segunda-feira, na direção da península de Yucatán, no México.
S2: Dean se transformou em um 'extremamente perigoso' furacão de categoria 4, a segunda mais alta da escala Saffir-Simpson.
S3: O Centro Nacional de Furacões dos Estados Unidos disse que ele pode ganhar força e chegar à categoria 5, potencialmente catastrófica, nas próximas 24 horas.
S4: Com ventos de cerca de 240 quilômetros por hora (km/h), o furacão estava movimentando-se ao sul das Ilhas Cayman às 5h (6h de Brasília) e seu centro estava a 185 quilômetros de distância da ilha principal, na direção oeste-nordeste.

Texto 3:
S1: Os brasileiros que estão em Cancún, um dos pontos turísticos mais visitados do México, estão se preparando para a chegada do furacão "Dean", que já matou nove pessoas na região do Caribe.
S2: Segundo o músico César Kiles, que está hospedado com sua mulher no Hotel Oasis, os turistas receberam orientação para se dirigirem para o abrigo subterrâneo do hotel caso o furacão passe pela região.
S3: "Soubemos do furacão anteontem, pela televisão. O hotel passou um informe avisando que talvez a gente tenha de ir para um abrigo que fica aqui embaixo", disse Kiles, em entrevista por telefone ao G1.
S4: Ele está em Cancún para passar a lua-de-mel com sua mulher, Aline, e disse que ainda não viu muitos indícios do furacão "Dean".
S5: "Por enquanto, o tempo está bom. Só o mar que está um pouco agitado, tem um pouco mais de onda", contou.
S6: César disse que ficará em Cancún até o próximo sábado e não está preocupado com a possibilidade de o furacão estragar sua viagem.
S7: "Não estou preocupado, não. Temos esperança de que tudo vai dar certo. O pessoal aqui está acostumado com furacões, então eles devem tirar isso de letra", concluiu o músico.
S8: O furacão "Dean" também está provocando transtornos para os brasileiros que vivem na Ilhas Caymans, que ficam entre a Jamaica e Cuba.

Figura 23: Três textos jornalísticos a serem sumarizados



Ciclo: S1.T3, S1.T2, S3.T1

Figura 24: Reações CST entre as sentenças dos textos da Figura 23

Primeiramente, o método verifica se o grafo das relações CST entre as sentenças possui ciclos. No exemplo, nota-se um ciclo entre as sentenças S1T2, S1T3 e S3T1. Avaliando-se o peso das relações presentes no ciclo, decide-se pela exclusão da relação Summary, existente entre S1T2 e S3T1. Feito isso, o algoritmo pode prosseguir com a ordenação topológica exata, já que o grafo não possui mais ciclos. A Figura 25 mostra um sumário extraído desses 3 textos, ordenado pelo método apresentado nesta seção. Esse sumário apresenta uma ordenação muito boa de suas sentenças, as quais apresentam as informações de maneira gradual e concisa.

Sumário:
S1T3: Os brasileiros que estão em Cancún, um dos pontos turísticos mais visitados do México, estão se preparando para a chegada do furacão "Dean", que já matou nove pessoas na região do Caribe.
S1T2: KINGSTON - O furacão Dean chegou à costa sul da Jamaica, inundando a capital e espalhando árvores e telhados depois de matar nove pessoas na passagem pelo Caribe nesta segunda-feira, na direção da península de Yucatán, no México.
S3T2: O Centro Nacional de Furacões dos Estados Unidos disse que ele pode ganhar força e chegar à categoria 5, potencialmente catastrófica, nas próximas 24 horas.
S5T1: O centro de Dean passou cerca de 160 km ao sul das ilhas Cayman, causando ventos de até 92 km/h.

Figura 25: Sumário Ordenado pelo método da ordenação topológica com análise semântica das relações CST

4. Avaliação e Resultados

4.1. Córpus e Medidas Utilizadas

Os trabalhos utilizam o córpus CSTNews (Cardoso et al, 2011), que possui atualmente 50 grupos de textos, sendo que cada grupo trata de um assunto diferente e tem em média três textos. Os textos foram coletados manualmente de jornais on-line por um período de 2 meses, entre Agosto e Setembro de 2007. As fontes dos textos foram os jornais on-line Folha de São Paulo, Estadão, O Globo, Jornal do Brasil e Gazeta do Povo. Essas fontes foram escolhidas devido a grande popularidade na web e também por trazerem as principais notícias do dia corrente, que é o que importa para o córpus, ou seja, uma mesma notícia publicada em fontes diferentes.

Para os testes dos métodos de ordenação de sentenças, foram utilizados sumários extraídos automaticamente pelo sumarizador CSTSumm (Jorge e Pardo,

2010), que é o melhor sumarizador multidocumento para o português desenvolvido até o momento.

Foi realizado um trabalho de ordenação manual de todos os 50 sumários gerados pelo CSTSumm, um para cada grupo do córpus CSTNews. Estes sumários ordenados manualmente serviram como sumários de referência para a avaliação dos sumários ordenados automaticamente pelos métodos deste trabalho. Essa tarefa de anotação do córpus foi feita seguindo-se critérios de coesão e apresentação das sentenças, de forma a proporcionar a melhor disposição delas no sumário.

O tempo necessário para a realização desta tarefa foi de aproximadamente 1 mês. A análise de informações redundantes entre as sentenças do sumário foi a maior dificuldade desta tarefa, pois, quando temos duas sentenças com o mesmo conteúdo, não temos informações suficientes para inferir alguma ordenação entre elas. A seguir, na Figura 26, temos um exemplo de um sumário produzido pelo sumarizador CSTSumm, com e sem a ordenação manual de suas sentenças.

Sumário com a ordenação de suas sentenças Na manhã desta sexta-feira o nadador Thiago Pereira conquistou sua quarta medalha de ouro pelos 200m medley na final da disputa no Complexo Aquático Maria Lenk, com a marca de 1m57s59. Em uma disputa emocionante, o Brasil conquistou nesta sexta-feira a medalha de ouro no revezamento 4x100 metros livre, uma das provas mais charmosas da natação, ao cravar o tempo de 3min15s90 (novo recorde pan-americano e sul-americano). O ouro no revezamento foi especial para Thiago Pereira (ele ganha a medalha mesmo não atuando na final), que conquistou sua segunda medalha dourada nesta sexta-feira e a quinta no Pan. O nadador conquistou mais dois ouros: um nos 200m medley e outro como reserva da equipe de revezamento 4x100m livre. Com as vitórias desta sexta, Thiago soma cinco ouros no Pan do Rio, tornando-se o novo recordista brasileiro em número de medalhas douradas em uma mesma competição.
Sumário sem a ordenação de suas sentenças: O nadador conquistou mais dois ouros: um nos 200m medley e outro como reserva da equipe de revezamento 4x100m livre. Na manhã desta sexta-feira o nadador Thiago Pereira conquistou sua quarta medalha de ouro pelos 200m medley na final da disputa no Complexo Aquático Maria Lenk, com a marca de 1m57s59. Em uma disputa emocionante, o Brasil conquistou nesta sexta-feira a medalha de ouro no revezamento 4x100 metros livre, uma das provas mais charmosas da natação, ao cravar o tempo de 3min15s90 (novo recorde pan-americano e sul-americano). O ouro no revezamento foi especial para Thiago Pereira (ele ganha a medalha mesmo não atuando na final), que conquistou sua segunda medalha dourada nesta sexta-feira e a quinta no Pan. Com as vitórias desta sexta, Thiago soma cinco ouros no Pan do Rio, tornando-se o novo recordista brasileiro em número de medalhas douradas em uma mesma competição.

Figura 26: Duas versões do mesmo sumário: com e sem a ordenação de suas sentenças

Para avaliar de maneira exata todos os métodos deste projeto, foram utilizadas duas medidas: o coeficiente de correlação de Spearman (Ts) e o coeficiente de correlação de Kendall (Tk) (Okazaki et al., 2004). O coeficiente de correlação de Spearman nos mostra o quanto as sentenças estão distantes de sua posição ideal, usando a distância euclidiana. O coeficiente de correlação de *Kendall* faz uma análise com relação a posição das sentenças usando a ordem ideal do sumário. Os dois resultados possuem 1 como melhor valor possível, quando as sentenças ordenadas pelo método estão exatamente na posição do sumário ordenado manualmente, e -1 como pior resultado, quando as sentenças estão totalmente invertidas, formando uma ordem decrescente.

Para calcular esses coeficientes, montamos a matriz de números inteiros π (2xN), onde cada elemento da primeira linha da matriz representa uma sentença do texto ordenado manualmente, e cada elemento da secunda linha representa a sentença ordenada por um dos dois métodos desenvolvidos. Por conveniência, definimos sempre o número de cada sentença baseando-se na sua posição no texto ordenado manualmente. Com isso, a primeira linha da matriz sempre será os números de 1 a N em ordem crescente. O exemplo a seguir ilustrará uma matriz π.

$$\pi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 2 & 4 & 1 \end{pmatrix}$$

A primeira linha da matriz apresenta as sentenças 1, 2, 3 e 4 na ordem em que estão dispostas no sumário ordenado manualmente. A segunda linha da matriz apresenta as sentenças 1, 2, 3 e 4 na ordem em que estão dispostas no sumário ordenado por algum método.

As equações dos coeficientes utilizados estão a seguir:

$$Ts = 1 - \frac{6}{N(N+1)(N-1)} \sum_{i=1}^{N} (\pi_{1i} - \pi_{2i})^2$$

$$Tk = \frac{1}{N(N-1)/2} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} sgn(\pi_{1j} - \pi_{1i}).sgn(\pi_{2j} - \pi_{2i})$$

onde Ts é o Coeficiente de Spearman, N é a quantidade de sentenças do sumario, π_{1i} é a posição da sentença i no sumário ordenado manualmente, π_{2i} é a posição da

sentença i no sumário ordenado pelo método a ser avaliado, sgn(x) = 1, para x > 0 e sgn(x) = -1, caso contrário.

4.2. Resultados e Conclusões

Os resultados de todos os métodos desenvolvidos neste projeto estão na Tabela 7, a seguir. São apresentados a média aritmética dos coeficientes, obtida com a execução dos algoritmos propostos, avaliados sobre os 50 sumários extraídos do corpus CSTNews.

Tabela 7: Resultados da avaliação dos métodos de ordenação.

Método	**Spearman**		**Kendall**	
	Média Aritmética	Desvio Padrão	Média Aritmética	Desvio Padrão
OPT – Método da posição Textual	**0,785**	0,284	**0,722**	0,304
OTSC – Método do tamanho (ordenação crescente)	0,186	0,558	0,135	0,482
OTSD – Método do tamanho (ordenação decrescente)	-0,186	0,558	-0,135	0,482
OTop -Ordenação Topical	**0,629**	0,379	**0,542**	0,363
OT - Ordenação Topológica	0,429	0,325	0,422	0,348
OT – Ordenação Topológica Invertida	0,523	0,393	0,444	0,348
OTCST – Ordenação Topológica com análise CST	0,511	0,350	0,490	0,386
OTCSTI – Ord. Topológica com análise CST Invertida	**0,603**	0,318	**0,516**	0,304

Nota-se que, mesmo sendo um método simples, o método de ordenação pela posição da sentença no texto fonte (OPT) se mostra mais eficaz que os demais. Também se nota que a análise das relações CST melhorou a qualidade do método que utiliza a ordenação topológica, porém, seu resultado ainda é inferior ao método da posição textual. A seguir, na Figura 27, temos duas versões do mesmo sumário, uma ordenada com o nosso melhor método (OPT) e a outra ordenada com o nosso pior método (OTSC). Nota-se uma grande diferença na legibilidade do sumário, onde o método OPT nos proporciona uma versão mais clara e concisa do sumário.

Sumário ordenado pelo método OPT
O CGE (Centro de Gerenciamento de Emergências) da Prefeitura de São Paulo registrava oito pontos de alagamento na cidade, às 9h30 desta segunda-feira.
A forte chuva em São Paulo complicava o trânsito na manhã desta segunda-feira, 16, e fez com que o Centro de Gerenciamento de Emergência (CGE) da Prefeitura colocasse a cidade em estado de atenção.
Naquele horário, segundo a CET (Companhia de Engenharia de Tráfego), havia 110 km de congestionamento em toda a cidade enquanto a média para o horário era de 76 km.

Sumário ordenado pelo método OTSD
A forte chuva em São Paulo complicava o trânsito na manhã desta segunda-feira, 16, e fez com que o Centro de Gerenciamento de Emergência (CGE) da Prefeitura colocasse a cidade em estado de atenção.
Naquele horário, segundo a CET (Companhia de Engenharia de Tráfego), havia 110 km de congestionamento em toda a cidade enquanto a média para o horário era de 76 km.
O CGE (Centro de Gerenciamento de Emergências) da Prefeitura de São Paulo registrava oito pontos de alagamento na cidade, às 9h30 desta segunda-feira.

Figura 27: Dois sumários ordenados pelos métodos OPT e OTSD

Como a esparsidade do grafo das relações CST é um fator que tem grande influência em nosso método, os métodos de ordenação topológica apresentam resultados inferiores aos demais. Outro fator que desencadeia esses resultados ruins é que os métodos de ordenação topológica ignoram completamente uma informação muito importante para essa tarefa de ordenação: a posição original da sentença no texto fonte.

Afirmando o que foi visto nos trabalhos de ordenação de sentença nos trabalhos correlatos apresentados na Seção 2.1, a informação topical realmente se mostrou uma informação muito importante para a produção de ordenações de sumários mais coerentes e coesos.

Como possibilidade para trabalhos futuros, tem-se a identificação semântica das relações topicais entre as sentenças, ou seja, uma análise mais profunda do significado dos tópicos existentes, servindo como base para novas heurísticas de ordenação de sentenças.

Referências

Aleixo, P. e Pardo, T.A.S. (2008). *CSTNews: Um Córpus de Textos Jornalísticos Anotados segundo a Teoria Discursiva Multidocumento CST (Cross-document Structure Theory).* Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, no. 326. São Carlos-SP, Maio, 12p

Barzilay, R.; Elhadad, M.; Mckeown, K. (2002). Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research,* Vol. 17, pp. 35-55.

Bollegala, D.; Okazaki, N.; Ishizuka, M. (2005). A machine learning approach to sentence ordering for multidocument summarization and it's evaluation. International Joint Conference on Natural Language Processing, *Lecture Notes in Artificial Intelligence,* Vol. 3651, pp. 624-635.

Chaves, A. R. and Rino, L.H.M (2008). The Mitkov Algorithm for Anaphora Resolution in Portuguese. In the *Proceedings of the International Conference on Computational Processing of Portuguese Language*, pp 51-60.

Jorge, M.L.C. and Pardo, T.A.S. (2010). Experiments with CST-based Multidocument Summarization. In the *Proceedings of the ACL Workshop TextGraphs-5: Graph-based Methods for Natural Language Processing*, pp. 74-82.

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In the *Proceedings of the Annual Meeting of ACL*, pp. 545-552.

Lin, C. and Hovy, E. (2001). NEATS: A multidocument summarizer. In the *Proceedings of the Document Understanding Conference (DUC01).*

Louis, A. and Nenkova, A. (2010). Creating Local Coherence: An Empirical Assessment. In the *Proceedings of the HLT '10 Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* pp 313-316.

Mani, I. (2001). *Automatic Summarization*. John Benjamins Publishing Co., Amsterdam.

Martins, C. B.; Pardo, T.A.S.; Espina, A.P.; Rino, L.H.M. (2001). *Introdução à sumarização automática.* Relatório Técnico RT-DC 002/2001. Departamento de Computação, Universidade Federal de São Carlos.

McKeown, K.R.; Barzilay, R.; Evans, D.; Hatzivassiloglou, V.; Kan M. Y.; Schiffman, B. and Tenfel, S. (2001). Columbia multi-document summarization: approach and evaluation. In the *Proceedings of the Document Understanding Conference (DUC01).*

Maziero, E.G.; Jorge, M.L.C.; Pardo, T.A.S. (2010). Identifying Multidocument Relations. In the *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science* - NLPCS, pp.60-69.

Okazaki, N.; Matsuo, Y.; Ishizuka, M. (2004). Improving Chronological Sentence Ordering by Precedence Relation. In the Proceedings of *COLING '04 Proceedings of the 20th international conference on Computational Linguistics,* pp 750-756.

Pardo, T.A.S. (2008). *Sumarização automática: principais conceitos e sistemas para o protuguês brasileiro.* Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional NILC - ICMC-USP (NILC-TR-08-04).

Salton, G.; Singhal A.; Mitra, M; Buckley C. (1997). Automatic Text Structuring And Summarization. *Information Processing & Management*, Vol. 33, No, 2, pp. 193-207.

Radev, D.R. (2000). A common theory of information fusion from multiple text sources, step one: Cross- ocument structure. In the Proceedings of the *1st ACL SIGDIAL Workshop on Discourse and Dialogue.*

Radev, D. R.; Jing, H.; Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In the *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*, Vol 4, pp 21-30.